D142–D144 *Nucleic Acids Research, 2004, Vol. 32, Database issue*
DOI: 10.1093/nar/gkh088

SMART 4.0: towards genomic data integration

Ivica Letunic, Richard R. Copley[1], Steffen Schmidt, Francesca D. Ciccarelli, Tobias Doerks, Jörg Schultz[2], Chris P. Ponting[3] and Peer Bork*

EMBL, Meyerhofstrasse 1, 69012 Heidelberg, Germany, [1]Wellcome Trust Centre for Human Genetics, Roosevelt Drive, Oxford OX3 7BN, UK, [2]Bioinformatik, Biozentrum, Am Hubland, University of Wuerzburg, 97074 Wuerzburg, Germany and [3]MRC Functional Genetics Unit, Department of Human Anatomy and Genetics, University of Oxford, South Parks Road, Oxford OX1 3QX, UK

Received September 17, 2003; Revised and Accepted October 6, 2003

ABSTRACT

SMART (Simple Modular Architecture Research Tool) is a web tool (http://smart.embl.de/) for the identification and annotation of protein domains, and provides a platform for the comparative study of complex domain architectures in genes and proteins. The January 2004 release of SMART contains 685 protein domains. New developments in SMART are centred on the integration of data from completed metazoan genomes. SMART now uses predicted proteins from complete genomes in its source sequence databases, and integrates these with predictions of orthology. New visualization tools have been developed to allow analysis of gene intron–exon structure within the context of protein domain structure, and to align these displays to provide schematic comparisons of orthologous genes, or multiple transcripts from the same gene. Other improvements include the ability to query SMART by Gene Ontology terms, improved structure database searching and batch retrieval of multiple entries.

INTRODUCTION

The SMART database (http://smart.embl.de; http://smart.ox.ac.uk) provides a tool to identify and annotate the signalling domains found in many eukaryotic proteins (1). The database consists of a library of Hidden Markov Models that are used to provide statistically robust inferences of the presence of specific domains in a particular sequence, and multiple sequence alignments of user query sequences with domains. The database provides extensive annotation for each domain, and is a comprehensive source of information on which proteins each is found in.

The primary motivation for the development of SMART was as a tool to study the evolution of function within multi-domain proteins. The availability of completed metazoan genomes, and increasing accuracy of prediction of gene structures and their multiple splice variants (2), has enabled us to create new extensions to SMART, allowing detailed overlaying of gene intron and exon structure with protein domain organization. This is coupled with a cross-referencing of orthologous genes in multiple genomes, collections of multiple splice variants of individual genes and new visualization tools to show schematic alignments of multiple gene structures. These new developments make SMART an ideal tool for studies of the evolution of gene and protein function.

PRE-CALCULATED RESULTS FOR COMPLETED METAZOAN GENOMES

In addition to the Swiss-Prot and spTrembl databases (3), which have been used by SMART since its inception, SMART's source sequence databases now include all available Ensembl proteomes (2). We compare sequences from all sources and generate a non-redundant set of proteins with multiple identifiers per sequence. Sequences are retrievable, and linkable, via any of the original identifiers.

IMPROVED DOMAIN COVERAGE

SMART continues to expand its domain coverage, with more than 70 new domains in the latest release, bringing the total close to 700. The rate of new, widespread domain discovery is falling, primarily as their numbers are limited (4). However, we continue to identify new domains of interest e.g. (5,6), and establish new links between others e.g. (7).

NEW DATABASE FEATURES

The core of SMART is a relational database management system (RDBMS) which stores information on SMART domains. In addition to previously available features (1,8), the SMART database now includes information on Pfam (9) domains in all proteins in NRDB. Users can now query the database for proteins that contain specific combinations of Pfam and SMART domains.

In addition to standard 'Domain selection' querying, it is now possible to find proteins based on Gene Ontology (GO) (10) terms associated with domains. Associations of domains with GO are taken from Interpro (11). GO querying is a two-step process. In the first step, the user obtains a list of domains matching the GO terms entered. After selecting the domains of interest from the list, proteins containing those domains are

*To whom correspondence should be addressed. Tel: +49 6221 387 526; Fax: +49 6221 387 517; Email: bork@embl-heidelberg.de

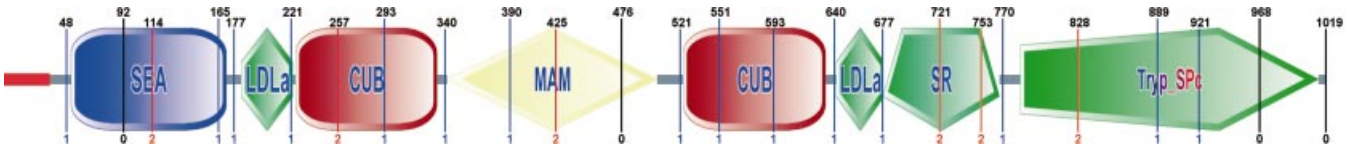

Figure 1. SMART representation of human enteropeptidase precursor (ENSP00000284885). Intron positions are represented by vertical lines, showing intron phase and exact position in AA. Intron positions are taken from Ensembl gene predictions.

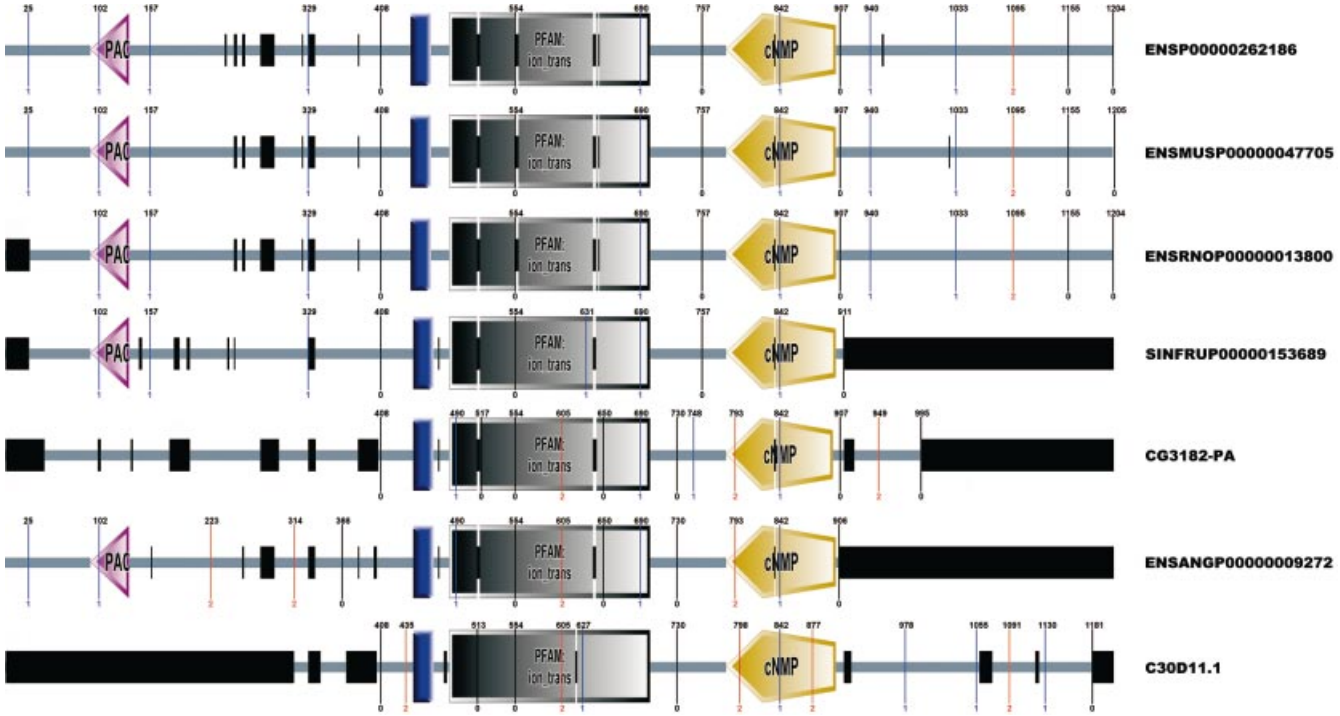

Figure 2. SMART representation of an orthologous group alignment. User-supplied proteins are aligned using ClustalW. Domains, intrinsic features and introns are mapped onto the alignment with their positions adjusted according to gaps (black boxes). This tool allows easy visual comparison of intron positions and their relations to protein features.

displayed. As with standard domain querying, results can be limited to specific taxonomic ranges.

SEQUENCE IDENTIFICATION

SMART uses the CRC64 algorithm to calculate checksums for all user-supplied sequences. If a matching checksum is found in the SMART database, pre-calculated results are displayed. Approximately 45% of all user-submitted sequences are identified in this way, resulting in shorter queues and much faster response times for all users.

Since user-supplied sequences can now be identified, several important new features have been introduced into SMART:

(i) Batch access: the SMART batch access facility allows users to submit multiple sequence identifiers or actual sequences, either by directly pasting the data into their web browser, or by uploading a file to the SMART server. If the user supplies plain sequences, their CRC checksums are calculated, and those with matches in the SMART database are displayed.

(ii) Intron positions shown in schematic protein figures: for proteins that match any of the Ensembl predictions, SMART will show intron positions as vertical coloured lines in graphical representations (Fig. 1). This information is retrieved from a pre-calculated mapping of Ensembl gene structures to protein sequences.

(iii) Extra information in the main results page: in cases where multiple IDs are associated with the same sequence, users get a list of all IDs with links to corresponding source databases. Since SMART now incorporates Ensembl genomes, users also get a list of alternative splices of the gene encoding the analysed protein (if there are any). It is possible to either display SMART protein annotation for any of the alternative splices, or get a graphical multiple sequence alignment of all of them.

IMPROVED SEARCHING OF STRUCTURE DATABASES

User sequences can now be searched against profiles derived from the SCOP database, using RPS-Blast (12,13). As well as detecting homologues of known structure, this enables easy

identification of the evolutionary superfamily to which any domains belong, and complements the links provided in domain annotation pages.

ORTHOLOGY INFORMATION

SMART provides orthology information for all Ensembl predicted proteins. These relationships are distinct from those provided by Ensembl. There are two separate sets of orthologues for each protein: 1:1 reciprocal best matches in other genomes and orthologous groups with reciprocal best hits from all genomes analysed (i.e. each of these proteins has exactly one orthologue in all six genomes). Orthologous groups are displayed as graphical multiple sequence alignments (Fig. 2). All orthology information is extracted from all-against-all Smith–Waterman (14) similarities for combined proteomes, using a previously described method (15).

CONCLUSIONS

With the growing number of completely sequenced eukaryotic genomes, the scientific community requires tools for easy comparative and large-scale analyses. With recent additions, we have expanded SMART's capabilities to accommodate the needs of many different types of user.

REFERENCES

1. Schultz,J., Milpetz,F., Bork,P. and Ponting,C.P. (1998) SMART, a Simple Modular Architecture Research Tool: identification of signaling domains. *Proc. Natl Acad. Sci. USA*, **95**, 5857–5864.
2. Clamp,M., Andrews,D., Barker,D., Bevan,P., Cameron,G., Chen,Y., Clark,L., Cox,T., Cuff,J., Curwen,V. *et al.* (2003) Ensembl 2002: accommodating comparative genomics. *Nucleic Acids Res.*, **31**, 38–42.
3. Bairoch,A. and Apweiler,R. (2000) The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res.*, **28**, 45–48.
4. Heger,A. and Holm,L. (2003) Exhaustive enumeration of protein domain families. *J. Mol. Biol.*, **328**, 749–767.
5. Ciccarelli,F.D., Bork,P. and Kerkhoff,E. (2003) The KIND module: a putative signalling domain evolved from the C lobe of the protein kinase fold. *Trends Biochem. Sci.*, **28**, 349–352.
6. Ciccarelli,F.D., Proukakis,C., Patel,H., Cross,H., Azam,S., Patton,M.A., Bork,P. and Crosby,A.H. (2003) The identification of a conserved domain in both spartin and spastin, mutated in hereditary spastic paraplegia. *Genomics*, **81**, 437–441.
7. Maurer-Stroh,S., Dickens,N.J., Hughes-Davies,L., Kouzarides,T., Eisenhaber,F. and Ponting,C.P. (2003) The Tudor domain 'Royal Family': Tudor, plant Agenet, Chromo, PWWP and MBT domains. *Trends Biochem. Sci.*, **28**, 69–74.
8. Letunic,I., Goodstadt,L., Dickens,N.J., Doerks,T., Schultz,J., Mott,R., Ciccarelli,F., Copley,R.R., Ponting,C.P. and Bork,P. (2002) Recent improvements to the SMART domain-based sequence annotation resource. *Nucleic Acids Res.*, **30**, 242–244.
9. Bateman,A., Birney,E., Cerruti,L., Durbin,R., Etwiller,L., Eddy,S.R., Griffiths-Jones,S., Howe,K.L., Marshall,M. and Sonnhammer,E.L. (2002) The Pfam protein families database. *Nucleic Acids Res.*, **30**, 276–280.
10. Ashburner,M., Ball,C.A., Blake,J.A., Botstein,D., Butler,H., Cherry,J.M., Davis,A.P., Dolinski,K., Dwight,S.S., Eppig,J.T. *et al.* (2000) Gene Ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature Genet.*, **25**, 25–29.
11. Mulder,N.J., Apweiler,R., Attwood,T.K., Bairoch,A., Barrell,D., Bateman,A., Binns,D., Biswas,M., Bradley,P., Bork,P. *et al.* (2003) The InterPro Database, 2003 brings increased coverage and new features. *Nucleic Acids Res.*, **31**, 315–318.
12. LoConte,L., Brenner,S.E., Hubbard,T.J., Chothia,C. and Murzin,A.G. (2002) SCOP database in 2002: refinements accommodate structural genomics. *Nucleic Acids Res.*, **30**, 264–267.
13. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
14. Smith,T.F. and Waterman,M.S. (1981) Identification of common molecular subsequences. *J. Mol. Biol.*, **147**, 195–197.
15. Zdobnov,E.M., von Mering,C., Letunic,I., Torrents,D., Suyama,M., Copley,R.R., Christophides,G.K., Thomasova,D., Holt,R.A., Subramanian,G.M. *et al.* (2002) Comparative genome and proteome analysis of *Anopheles gambiae* and *Drosophila melanogaster*. *Science*, **298**, 149–159.